

Contents

1	Introduction of Video Analytics	1
1.1	Procedures of Video Analytics	1
1.2	Key Performance Metrics	2
1.3	Applications of Video Analytics	3
1.4	Summary	5
1.5	Problems	6
2	Fundamentals of Edge Video Analytics	7
2.1	Concept and Paradigms	7
2.1.1	Concept of Edge Computing	7
2.1.2	Paradigms of Edge Computing	10
2.2	Edge Video Analytics	15
2.2.1	Definition of Edge Video Analytics	15
2.2.2	Architectures of Edge Video Analytics	16
2.2.3	Challenges in Edge Computing for Video Analytics	19
2.3	Summary	20
2.4	Problems	21
3	Fundamentals of Artificial Intelligence for Edge Video Analytics	23
3.1	Historical Development	23
3.1.1	The Birth of Computer Vision	23
3.1.2	The Enlightenment Period	24
3.1.3	The Exploration Period	24
3.1.4	The Flourishing Period	25
3.2	Traditional Computer Vision Methods	25
3.2.1	Scale-Invariant Feature Transform	26
3.2.2	Histogram of Oriented Gradients	28
3.3	Traditional Artificial Neural Networks	30
3.3.1	McCulloch-Pitts Neuron Model	30
3.3.2	Perceptron Model	34
3.3.3	Multilayer Perceptron Model	36
3.3.4	Backpropagation Algorithm	37

3.4	Convolutional Neural Networks	41
3.4.1	Architectures of CNNs	41
3.4.2	Image Classification	44
3.4.3	Object Detection	49
3.5	Generative Adversarial Networks	55
3.5.1	Architectures of Generative Adversarial Networks	56
3.5.2	Objective Function and Training of Generative Adversarial Networks	57
3.5.3	Application of Generative Adversarial Networks	59
3.6	Transformers	64
3.6.1	Architectures of Transformer	65
3.6.2	Application of Transformer	70
3.7	Summary	73
3.8	Problems	74
4	Model Training for Edge Video Analytics	77
4.1	Training Architecture for Edge Video Analytics	77
4.1.1	Training at the Cloud	77
4.1.2	Training at the Edge	79
4.1.3	Training Both at the Cloud and at the Edge	84
4.2	Training Method for Edge Video Analytics	86
4.2.1	Federated Learning	86
4.2.2	Transfer Learning	102
4.2.3	Mixture of Experts	109
4.3	Summary	113
4.4	Problems	114
5	Model Deployment for Edge Video Analytics	115
5.1	Quantization	115
5.1.1	Quantization Function	116
5.1.2	Post-Training Quantization	119
5.1.3	Quantization-Aware Training	120
5.2	Network Pruning	122
5.2.1	Pruning Granularity	124
5.2.2	Neural Network Pruning Algorithms	124
5.3	Knowledge Distillation	127
5.3.1	Offline Distillation	128
5.3.2	Online Distillation	129
5.3.3	Self-Distillation	131
5.3.4	Data-Free Distillation	132
5.4	Lightweight Model Design	133
5.4.1	Factoring Convolution	133
5.4.2	Separating and Grouping Convolution	136
5.5	Summary	140
5.6	Problems	142

6	End-Edge-Cloud Orchestration for Video Analytics	143
6.1	Knob Configuration Adaptation	145
6.1.1	Statistical Model-Based Methods	146
6.1.2	Measurement-Based Methods	147
6.1.3	Deep Learning-Based Methods	148
6.2	Video Temporal and Spatial Pruning	150
6.2.1	Video Temporal Pruning	150
6.2.2	Video Spatial Pruning	151
6.2.3	Amalgamating Both Temporal and Spatial Pruning	153
6.3	NN Partition Among End, Edge, and Cloud	153
6.3.1	Layer-Wise Profiling-Based Methods	154
6.3.2	DAG-Based Methods	155
6.3.3	Online Algorithms	156
6.3.4	Model Early Exit	157
6.4	Multi-Edge Query Scheduling	158
6.4.1	Centralized Scheduling	159
6.4.2	Distributed Scheduling	160
6.5	Summary	161
6.6	Problems	163
7	Platforms for Edge Video Analytics	165
7.1	Practical Deployment of Edge VA	165
7.1.1	Hardware and Software for Edge VA Deployment	165
7.1.2	Open-Source Datasets for Edge VA	167
7.1.3	Commercial Platforms for Edge VA Deployment	167
7.1.4	Industrial Deployments of Edge VA	168
7.2	Research Efforts on Edge VA Platforms	168
7.2.1	Edge VA Platforms for Edge Devices	169
7.2.2	Edge VA Platforms for Local Area	169
7.2.3	Edge VA Platforms for Wide Area	170
7.3	Summary	171
7.4	Problems	171
	References	173
	Index	189

MLP	Mean Average Precision
MMR	Mobile Augmented Reality
MRU	Magnetic Resonance Imaging
MVS	Multi-View Stereo
NAS	Neural Architecture Search
NFV	Network Functions Virtualization
NPU	Neural-network Processing Unit
PET	Positron Emission Tomography
QoS	Quality of Service
RBC	Red Blue Green